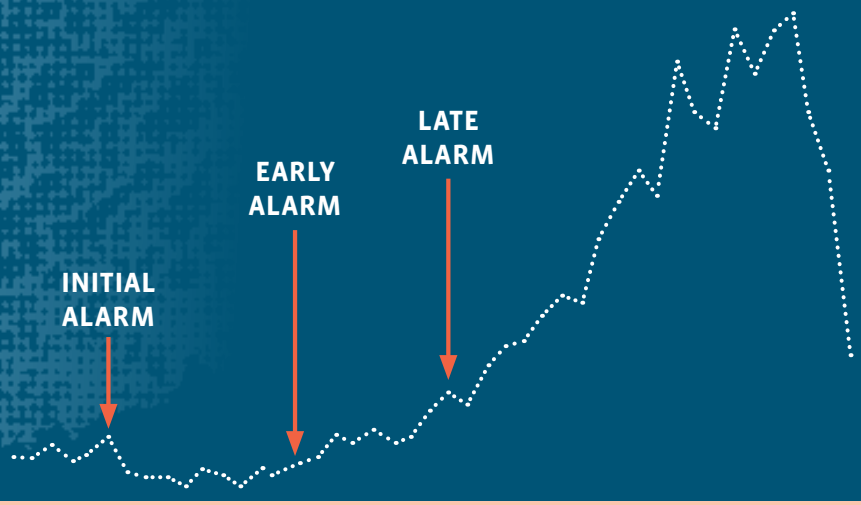
OPERATIONAL GUIDE



Early Warning and Response System (EWARS) for Dengue Outbreaks

World Health Organization
TDR For research on diseases of poverty
UNICEF • UNDP • World Bank • WHO

WHO Library Cataloguing-in-Publication Data

Operational guide: Early Warning and Response System (EWARS) for dengue outbreaks

ISBN 978-92-4-151205-3

© World Health Organization 2017

Some rights reserved. This work is available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; https://creativecommons.org/licenses/by-nc-sa/3.0/igo).

Under the terms of this licence, you may copy, redistribute and adapt the work for non-commercial purposes, provided the work is appropriately cited, as indicated below. In any use of this work, there should be no suggestion that WHO endorses any specific organization, products or services. The use of the WHO logo is not permitted. If you adapt the work, then you must license your work under the same or equivalent Creative Commons licence. If you create a translation of this work, you should add the following disclaimer along with the suggested citation: "This translation was not created by the World Health Organization (WHO). WHO is not responsible for the content or accuracy of this translation. The original English edition shall be the binding and authentic edition".

Any mediation relating to disputes arising under the licence shall be conducted in accordance with the mediation rules of the World Intellectual Property Organization.

Suggested citation. Operational guide: Early Warning and Response System (EWARS) for dengue outbreaks. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Sales, rights and licensing. To purchase WHO publications, see http://apps.who.int/bookorders. To submit requests for commercial use and queries on rights and licensing, see http://www.who.int/about/licensing.

Third-party materials. If you wish to reuse material from this work that is attributed to a third party, such as tables, figures or images, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.

General disclaimers. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of WHO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers' products does not imply that they are endorsed or recommended by WHO in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by WHO to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall WHO be liable for damages arising from its use.

Graphic Design: Lisa Schwarb, Lausanne
Printed by the WHO Document Production Services (DUP), Geneva, Switzerland

OPERATIONAL GUIDE

Early Warning and Response System (EWARS) for Dengue Outbreaks

World Health Organization

TDR For research on diseases of poverty
UNICEF • UNDP • World Bank • WHO

Contents

Foreword vi
Preface vii
Acknowledgements viii
Glossary ix

Chapter 1. Preparing your dataset 1

1.1 General requirements for the dataset 1
1.2 Stata packages installation, prior to the analysis 1
1.3 Datasets, do-files and folders 2
1.4 Surveillance data, contents, format and location 4
1.4.1 General variables 4
1.4.2 Outbreak indicator 4
1.4.3 Alarm indicator(s) 5

Chapter 2. Data management and retrospective analysis 7

2.1 Calibrating the do-file 7
2.1.1 Three easy steps before you begin 7
2.1.2 Changing the general settings in the user's do-file 8
2.1.3 Adjusting the settings for the outbreak indicator 11
2.1.4 Adjusting the settings for the alarm indicator(s) in the user's do-file 12
2.1.5 Settings for running graphics 14
2.1.6 Advanced settings 15
2.1.7 Run the do-file 19
2.1.8 Interpreting the results of the prospective data 20
2.1.9 Interpreting the results from an operational perspective 21

Chapter 3. Using the early warning and response system 23

3.1 Prospective early warning system 23
3.1.1 Data entry 24
3.1.2 Interpretation of staged response 24

ANNEX 1. Technical guide 27

1. Introduction 27
2. Methodology 27
   2.1 Retrospective phase 27
   2.2 Prospective early warning phase (i.e. surveillance workbook Excel file) 28
3. Structural design (Stata and Excel) 28
   3.1 Phase I (retrospective) 28
   3.2 Phase II (prospective surveillance workbook) 34
4. Potential errors 35
   4.1 Problem related to Step 2: Few Data (error code r 2000) 35
   4.2 Problem related to Step 7: Convergence not achieved (error code r 430) 35
   4.3 Problem related to Step 8: too low/ high values of "season length" or "outbreak period" (error code r 2000) 36
   4.4 Problem related to Step 16: open/ active excel file before running the analysis (error code r 603) 36
   4.5 Problem related to Step 8: too small sample size (error code r 198) 37
   4.6 General error: wrong district code (error code r 111) 37

LIST OF FIGURES

Fig. 1. Installing and Updating Software Packages 2
Fig. 2. Cleaning Your Dataset 3
Fig. 3. Summary List of Variables 4
Fig. 4. Where to Start? 8
Fig. 5. General Settings 9
Fig. 6. Calibrating the Outbreak Indicator 11
Fig. 7. Calibrating the Alarm Indicator 12
Fig. 8. Graphical Settings 14
Fig. 9. Advanced Settings 15
Fig. 10. Modelling the Endemic Channel 16
Fig. 11. Run the Do-File. 19
Fig. 12. Absolute Values in the Output Table 20
Fig. 13. Sensitivity and PPV 21
Fig. 14. Prospective Data Entry and Interpretation 23
Fig. 15. Results Interpretation: An Illustrated Example 25
Fig. 16. Block Diagram of the Retrospective/ Prospective Phases 30
Fig. 17. Illustrated Process of Alarm Window Size 31
Fig. 18. Illustration of Prediction Distances 32
Fig. 19. Example Error Message: “No Observations” 35
Fig. 20. Example Error Message: “Convergence not Achieved” 35
Fig. 21. Example Error Message: “r2000” 36
Fig. 22. Example Error Message: “File [...] Not Saved” 36
Fig. 23. Example Error Message: “Sample Size Too Small” 37
Fig. 24. Example Error Message: “Variable_000001 Not Found” 37

FOREWORD

Dengue outbreaks threaten the stability of national health systems worldwide. Every year, late detection and inadequate response mechanisms compound the effects of rapid dengue transmission. Although outbreak alert indicators exist, the means to deploy them in early warning systems is often lacking.

Within this context, a programme led by TDR, the Special Programme for Research and Training in Tropical Diseases, conducted multi-country research into alarm signals for outbreaks and their use within early warning systems. In line with the prevailing literature,[1,2,3] alarm variables such as probable dengue cases, mean temperature (and, to a lesser extent, rainfall and humidity) all evidenced predictive abilities.[4] However, it was clear that countries lacked the requisite skills and resources to deploy these alarm signals in a predictive, operational way. It was on this basis that an accessible, adaptable and user-friendly early warning system was developed.

The aim of the *Early Warning and Response System (EWARS) for Dengue Outbreaks:* operational guide is to provide programme managers with a user-friendly tool that can: (i) analyse and draw conclusions from historic dengue datasets; (ii) identify appropriate alarm indicators that can sensitively and specifically predict forthcoming outbreaks at smaller spatial scales; and (iii) use these results and analyses to predict and build an early warning system to detect dengue outbreaks in real-time. Together, these three components build technical capacity and provide a standardized methodology for predicting dengue outbreaks in countries where skills and resources are currently constrained.

This guide was produced by TDR together with the World Health Organization's Neglected Tropical Diseases (WHO/NTD) and WHO regional offices in the context of a European Union-financed research programme, the International Research Consortium on Dengue Risk Assessment, Management and Surveillance (IDAMS), to develop an evidence-based, early warning system for outbreak detection and management of dengue fever outbreaks.

---

1 Hii YL, Zhu H, Ng N, Ng LC, Rocklöv J (2012). Forecast of dengue incidence using temperature and rainfall. *PLoS Negl Trop Dis.* 6:e1908.

2 Halide H, Ridd P (2008). A predictive model for dengue hemorrhagic fever epidemics. Int J Environ Health Res. 18:253–265.

3 Hii YL, Rocklöv J, Wall S, Ng LC, Tang CS, Ng N (2012). Optimal lead time for dengue forecast. *PLoS Negl Trop Dis.* 6:e1848.

4 Bowman LR, Tejeda GS, Coelho GE, Sulaiman LH, Gil BS, McCall PJ et al. (2016). Alarm variables for Dengue Outbreaks: a multi-centre study in Asia and Latin America. PLoS One. 11:e0157971.

PREFACE

Welcome to this *Operational Guide to the Early Warning and Response System (EWARS) for Dengue Outbreaks*. This guide will provide you with the necessary information and tools to use and analyze surveillance data to predict dengue outbreaks.

Step-by-step instructions are detailed in the chapters below to help you: i) organize your raw data; ii) enter it into computer software; iii) run the computer software; and iv) interpret your results. Each of the steps is illustrated for better clarity and complemented by a troubleshooting section found at the end of the guide.

A technical Annex provides information on the processes and statistics used to construct the prediction model, although this will be unnecessary for all but a few advanced users who want to understand the original source code.

Finally, it is most important that the data you have collected is in a format recognizable by the statistical software package, Stata*. The first chapter walks you through a process of data cleaning and transformation to ensure that this will be the case. Good luck!

* StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP..

ACKNOWLEDGEMENTS

The *Early warning and Response System (EWARS) for Dengue Outbreaks:* operational guide was written by Laith Hussain-Alkhateeb, Leigh Bowman and Max Petzold. It was coordinated and supported by Axel Kroeger and Piero Olliaro of TDR, the Special Programme for Research and Training in Tropical Diseases. Special thanks goes to Ali F. Majeed for providing insight and contributions to the development of the programming materials. We also acknowledge David Benitez (Mexico) and Balvinder Gill (Malaysia) for beta testing the materials, as well as the following colleagues for contributing valuable country data to enable the development of EWARS: Roberta Carvalho (Brazil); Giovanini Coelho (Brazil); Lokman Hakim (Malaysia); Luong Quang (Viet Nam); Ronald Ramm (Dominican Republic); and Gustavo Tejeda (Mexico). This guide was developed in the context of a dengue research programme supported by a grant from the European Commission (Grant Number m281803) to the International Research Consortium on Dengue Risk Assessment, Management and Surveillance (IDAMS) within the 7th Framework Programme of the European Commission (FP7), and by TDR.

GLOSSARY

Term	Definition
Stata	Software used to process and analyse your data and to generate meaningful results.
Do-file	This is part of the software package, a platform (sub-screen or sheet) which allows the user to write commands and instruct Stata on the order in which to process commands.
User interface.do	This is one of two do-files that are part of the package. This is the do-file that the user will be using to change some general settings before running the software.
Dengue_Analysis.do	This is the second do-file from the package. This file is ONLY to be saved in the same folder where the other materials (i.e. the user's interface do-file, dataset and the surveillance workbook excel file) are saved and is not for the user's use. This file is linked to the user-interface.do file and contains all of the code relevant for building the early warning system in Stata.
Surveillance workbook	This is an Excel file provided with the package. This file should be located in the same folder as the other files. This file is used for the prospective early warning system analysis.
Endemic channel	This represents the number of cases within the expected normal seasonal range; anything above this moving threshold would be considered representative of an unprecedented number of cases, i.e. an outbreak.
Outbreak indicator	The dependent variable(s) used to define outbreaks. Usually probable or hospitalised dengue cases.
Alarm indicator	The independent variable(s) used to predict outbreaks. This could be one of a number of meteorological variables e.g. rainfall, temperature, or other entomological/ epidemiological variables.
Spline	A function to capture both positive and inverse associations between the same alarm indicator(s) and outbreak indicator(s).
Sensitivity	The proportion of alarms that successfully predicted outbreaks. A higher number indicates higher true positive alarms.
Positive predictive value (PPV)	The proportion of alarms that unsuccessfully predicted outbreaks. A higher number indicates lower false positive alarms.
SD	This is the standard deviation, which is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean (average).

Chapter 1
Preparing your dataset

1.1 General requirements for the dataset

To begin, your dataset needs to be organized in a way that Stata can understand it. Follow the points below to ensure your surveillance dataset will be compatible and to avoid error messages.

Stata is a statistical program that extracts data from your surveillance dataset, which may be contained in an Excel file, for example. It is vital that your dataset, including column headings and row names, follow the rules below.

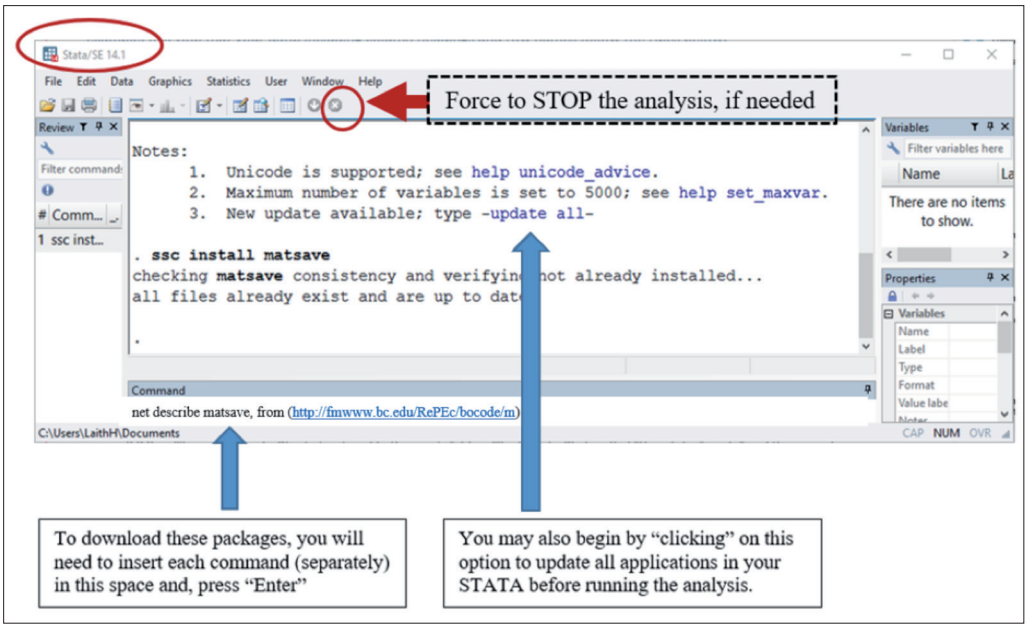
- All variable names in both your surveillance dataset and Stata need to strictly follow the same spelling, letter case and spaces, as written in the surveillance dataset.
- There should be no additional symbols/characters entered into the surveillance data, for example "_", "-", "/", ";", "@", etc.
- No row names are permitted.
- Missing data (observations) in the Excel file should be kept as "empty" cells, i.e. not "0".
- All variable data should be stored in a numeric format (no text or "string" data). Column headings should be the first row of the dataset.
- All data should be collected/reported on a weekly basis and at the same spatial scale, usually a small administrative unit, such as the district or locality.

1.2 Stata packages installation, prior to the analysis

You will need to install Stata as well as additional 'packages' before running any analyses.

- Stata 14.1 version was used in this analysis.
- After installing this version (Stata 14.1), you will need to install "mkmat" and "matsave" packages. To install these packages; for "matsave", simply type "net describe matsave", from (http://fmwww.bc.edu/RePEc/bocode/m) – then press "Enter" followed by "net install matsave.pkg" then press "Enter", in the Stata program interface.
- For "mkmat" package, type - net describe ip6, from (http://www.stata.com/stb/stb20) – then press "Enter" followed by "net install ip6.pkg" then press "Enter". See Figure 1 for further illustrations.

Figure 1. Installing and Updating Software Packages



1.3 Datasets, do-files and folders

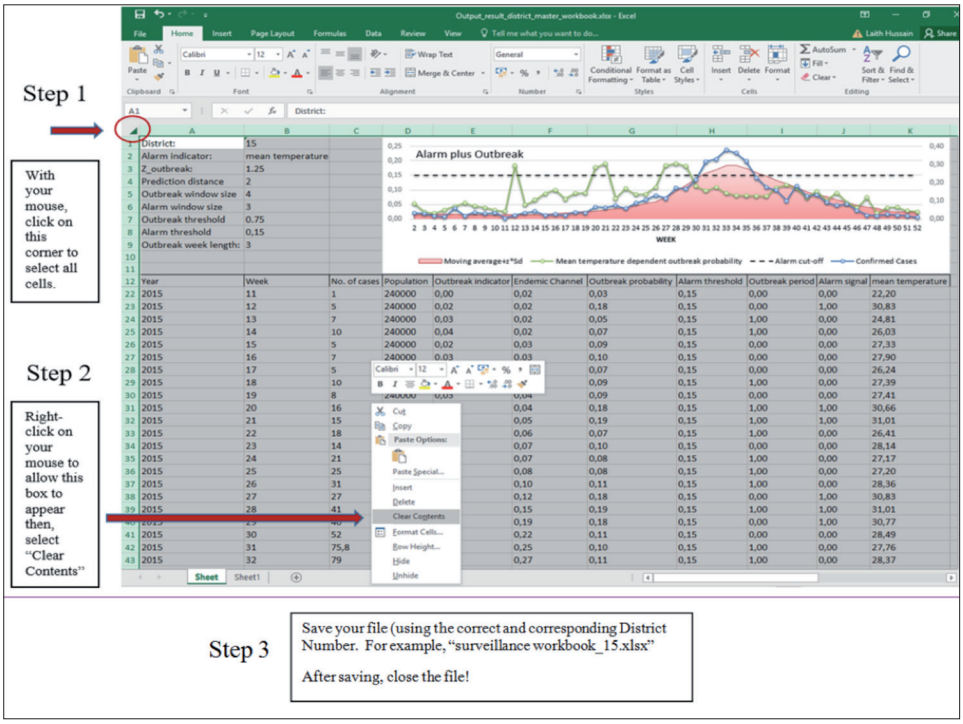
A do-file can be defined as: the user's platform where you can enter instructions, commands or additional comments to run the model in Stata. The entire package comprises four files that need to be stored in the same folder on your computer. Please make sure that you have two do-files, one surveillance dataset and one surveillance workbook in this folder. The points below describe other important aspects of the package.

There are two do-files. The first do-file is called "User_interface.do" and the second, "Dengue_Analysis.do". The user should only use the "User_interface.do" do-file. Do not edit "Dengue_Analysis.do".

- Your surveillance dataset must be stored as an Excel spreadsheet (.xlsx).
- The surveillance workbook is called "surveillance workbook_DistrictNumber". This file should be saved with the extension ".xlsx".
-

- The “surveillance workbook_DistrictNumber.xlsx” should not be open during the analysis. The user can open the “surveillance workbook_DistrictNumber.xlsx” once the analysis process under the “User_interface.do” do-file is complete. Stata will tell you when the model has finished running.
- The surveillance workbook Excel file should not have any data in it. Figure 2 illustrates how to delete old data.

Figure 2. Cleaning Your Dataset



1.4 Surveillance data, contents, format and location

Before running the analysis, you will need the list of variables described below. At least one alarm indicator is required (e.g. mean temperature) but additional alarm indicators can also be included (Figure 3).

Figure 3. Summary List of Variablesa

	A	B	C	D	E	F	G	H
1	year	week	week qweekly_hospitalised_cases	humidety_weekly_mean	weekly_rain_sum	temp_weekly_mean	population_annual_number	district
2	2015	2	3	78,3	0	24,7	200000	1
3	2015	3	0	85,9	8,1	25,7	200000	1
4	2015	4	1	85,1	84,3	23,4	200000	1
5	2015	5	1	79,1	0	20,5	200000	1
6	2015	6	1	81,9	35,3	24,5	200000	1
7	2015	7	3	71,4	6,3	21,6	200000	1
8	2015	8	1	80,0	47,3	25,7	200000	1
9	2015	9	2	79,1	6,6	24,6	200000	1
10	2015	10	4	77,0	1	21,8	200000	1
11	2015	11	0	71,7	0,3	25,6	200000	1
12	2015	12	2	74,9	0,5	24,6	200000	1
13	2015	13	1	82,1	68,3	23,9	200000	1
14	2015	14	0	79,1	2,8	24,8	200000	1
15	2015	15	4	86,7	0	25,5	200000	1
16	2015	16	2	74,6	1	24,9	200000	1
17	2015	17	1	83,7	0	23,6	200000	1
18	2015	18	1	77,7	0	25,7	200000	1
19	2015	19	0	81,0	0	24,6	200000	1
20	2015	20	0	81,7	0	21,8	200000	1
21	2015	21	4	82,9	0,9	25,6	200000	1
22	2015	22	0	82,7	72,7	24,6	200000	1
23	2015	23	1	86,6	8,4	23,9	200000	1
24	2015	24	0	85,4	6,3	24,8	200000	1
25	2015	25	0	82,1	2,3	25,5	200000	1
26	2015	26	0	84,1	11,2	24,9	200000	1
27	2015	27	0	88,7	7,8	23,6	200000	1
28	2015	28		86,3	55,8	25,7	200000	1
29	2015	29	0	86,7	16,8	24,6	200000	1
30	2015	30	0	85,3	1,5	21,8	200000	1

The user can add more "alarm" indicators here!

1.4.1 General variables

- "year": the year when data were captured. The year must be entered in full using "four" numbers (e.g. 2015, 2016, etc.).
- "week": the number of the epidemiological week (Sunday to Saturday) when data captured. The week number must be entered in full (e.g. 1, 2, 3…etc.).
- "district": a number (code) that represents the district where data were captured (e.g. 1, 2, 3, 15, 22, etc.).
- "population": the annual population size of a district is reported in absolute numbers in the surveillance data. See column "G" in figure 3, above.

1.4.2 Outbreak indicator

"weekly_hospitalised_cases": this is the number of hospitalized cases in a given district per epidemiological week. As a substitute, you could enter probable case data here, but this must be consistent across all datasets. You cannot mix hospitalized and probable case data.

1.4.3 Alarm indicator(s)

You need a minimum of one alarm indicator, but there is no maximum limit. Please ensure that your alarm indicator is labelled in the column as shown in the example below:

- "meanhumidity": the weekly mean humidity (as a percentage), for a corresponding district and year;
- "weeklyrainfall": the total weekly rainfall (in cm) for a corresponding district and year;
- "meantemperature": the weekly mean temperature in either Celsius or Fahrenheit (do not combine Celsius and Fahrenheit data in the same spreadsheet: choose one or the other), for a corresponding district and year.

Please note that the user is given the option (in the user's interface do-file) to insert the corresponding variable names as they appear in the surveillance data (column). It is important though that the corresponding variable name is entered correctly where specified in the do-file, i.e. should have an exact spelling and format.

Chapter 2
Data management and retrospective analysis

You will be given two do-files; the "user_interface.do" and the "Dengue_Analysis.do" to save in your folder, as mentioned earlier. However, you are ONLY going to use the "user_interface.do" when making all necessary settings and calibrations. The second, "Dengue_Analysis.do" do-file, is an automated-platform to respond to and process your defined settings that will take place in the "user_interface.do". Therefore, it is necessary to save both do-files in the same folder to ensure they are both linked to each other.

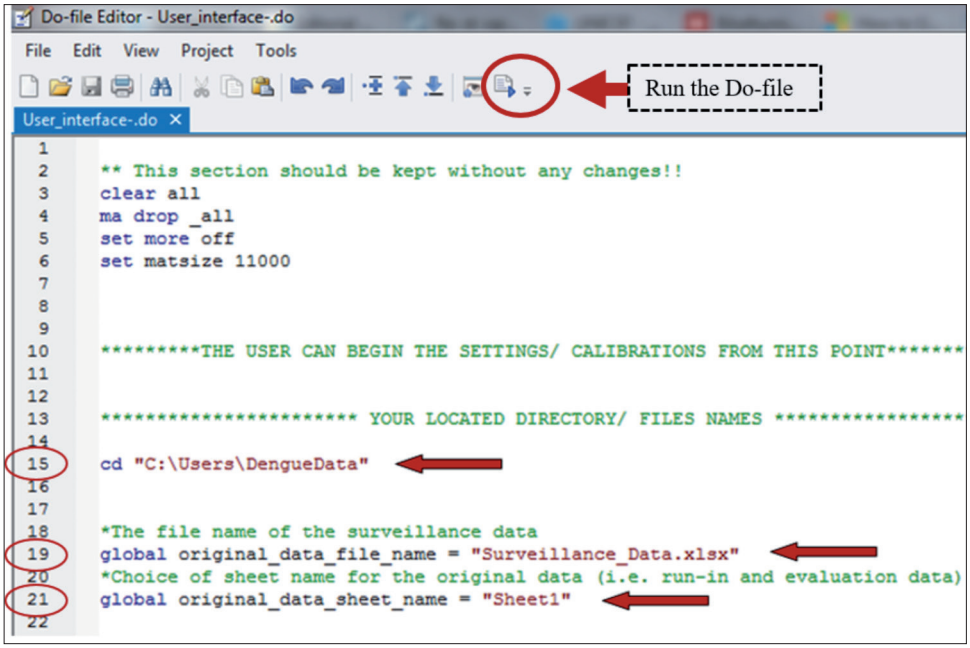
The instructions below relate to the first do-file (user_interface.do). You do not need to open this file.

2.1 Calibrating the do-file

2.1.1 Three easy steps before you begin

- Before running the do-file, you must change the model settings in accordance with your data and the analysis you want to run (changing values, alarm/outbreak thresholds, etc.). This will improve the prediction of outbreaks.
- Figures 4-9 (in next pages) show you where to find these settings so that you can change them in your do-file.
- In the do-file, each line of code has a line number. To help you follow these instructions, you can refer to the line number to see where you need to make the change in your do file.

Figure 4. Where to Start?



Line number: 15

Choosing where your results and outputs will be saved

cd "C:\Users\DengueData"

- Create a folder on your desktop. Name it "Stata Analysis". Copy the pathway of the folder (usually you can right click, click properties and you will see the pathway, it looks like this: C: Desktop\User). Copy this pathway and enter it in between the quotation marks in the do-file, as shown above. Now all your results and analyses will be stored in this folder.

Line number: 19

Finding your surveillance dataset

```
global original_data_file_name = "Surveillance_Data.xls"
```

- Find your dataset on your computer. Then copy the full name of your dataset file including the extension (.xls or xlsx) and paste it between the quotation marks in your do-file, as shown above.

Line number: 21

Identify the name of the data "sheet" where the records are located

```
global original_data_file_name = "Sheet1"
```

- In Excel, a spreadsheet is divided into "sheets". These can be given a name, but unless you do, the usual name is sheet1. Enter (copy/paste) into the do-file the EXACT "sheet" name (text) and number (e.g. "Sheet1") between the quotation marks so Stata can find your dataset.
- *This precedes* **step1** *in the annex.*

2.1.2 Changing the general settings in the user's do-file

Figure 5. General Settings

Note: more settings are illustrated in the section below.

```
24  ********************** GENERAL SETTINGS **********************
25
26
27  **Choice of run-in period
28  *Specify the year/week where the run-in period stops (i.e. evaluation period st
29  scalar stop_runin=201252
30
31
32  *Export_results=0; estimating the parameters from the run-in data and performs
33  *Export_results=1; estimating the parameters for of the entire data (both run-i
34  scalar generating_surveillance_workbook=1
35
36
37  *Perfroms the analysis per specific distric(s) (i.e. specifying the correspondi
38  global run_per_district 15
39
```

Line number: 29

Dividing your dataset

```
scalar stop_run-in = 201252
```

- To detect alarm indicators that can help you predict outbreaks, you need to first run a retrospective analysis of your data. For this step, the model requires that you divide your dataset into two time periods. Now you must choose the dates of your historic/run-in period and the evaluation period.
- To do so, enter the year and week that you want your run-in period to END. Stata will automatically use all data after this date as the evaluation period. For example, enter **201326** with no space between the year and week digits for year "2013" and week "26".
- *This corresponds to* **steps 1** *and* **10** *in the annex.*

Line number: 34

Retrospective analysis or prospective early warning

Here you can choose between a retrospective analysis as defined above or, if you have already run the retrospective analysis and have found your alarm indicators, you can build your prospective early warning system.

```
scalar generating_surveillance_workbook= 0
```

- Here you simply have to select '0' or '1'.
- Choosing result=0 runs a retrospective analysis of your dataset that allows you to find your alarm indicators. You should do this first (choosing result=0) every time you make a change to the settings in the do-file.
- Choosing result=1 generates the prospective early warning system in Excel. Only choose this if you have already run the retrospective analysis, made all necessary calibrations and have already identified your alarm indicators.
- *This corresponds to* **step 16** *in the Annex.*

Line number: 38

Choosing districts to be analysed

```
global run_per_district 15 17
```

- Here you can choose to analyse specific districts or all districts.
- Enter the desired district code(s) of interest. By leaving this option "empty", Stata will analyse all districts.
- The user can enter more than one district code by leaving one space between each district code.
- *This precedes* **step 1** *in the annex.*

2.1.3 Adjusting the settings for the outbreak indicator

Figure 6. Calibrating the Outbreak Indicator

```
42  ************************ OUTBREAK-RELATED SETTINGS ****************************
43
44  *Choice of the variable names as written in the surveillance data for each of:
45  *population: the name of variable which represents the "the total annual number
46  global population = "population"
47  *number of cases: the name of variable which represents the number of cases (i.e
48  global number_of_cases = "weekly_hospitalised_cases"
49
50  *Choice of the number of outbreak weeks to declare outbreak period/stop outbreak
51  *For example, if outbreak_week_length=3 then minimum of "3" outbreak weeks in a
52  scalar outbreak_week_length=3
53
```

Here we describe 3 steps that you need to alter.

Line number: 46

Defining the annual number of people in your districts

global population = "population"

- In each district you will have a different human population.
- Here, you must type the name of the variable that tells Stata the size of each district population, e.g. if your spreadsheet column is labelled 'population', please type 'population' into the do-file.
- It is important to type the EXACT text of the variable names (i.e. the title of the column) as it appears in the surveillance data.
- *This corresponds to* **step 3a** *in the annex.*

Line number: 48

Labelling the number of cases

global number_of_cases = "weekly_hospitalised_cases"

- Here you must consider what incident case data have been captured. For example, it might be the number of weekly hospitalized cases (recommended) or weekly probable clinical cases.
- Please type between the quotation marks the column name that describes these data – it must be the same as in your surveillance dataset.
- *This corresponds to* **step 3a** *in the annex.*

Line number: 52

Defining the outbreak period

scalar outbreak_period_length = 3

- A collection of consecutive outbreak weeks defines an outbreak. It is up to you how to define an outbreak, but we recommend 2 weeks.
- Enter the desired choice of the minimum number of outbreak weeks needed to define the outbreak period.
- For example, if outbreak_week_length=3 then a minimum of "3" consecutive outbreak weeks is required to define an outbreak (outbreak period).
- *This corresponds to* **step 3b** *in the annex.*

2.1.4 Adjusting the settings for the alarm indicator(s) in the user's do-file

Figure 7. Calibrating the Alarm Indicator

```
55  ****************************** ALARM INDICATOR-RELATED SETTINGS ********************
56
57  *Choice of alarm indicator: choice of N alarm indicators to run models of multiple or single alarm indicator.
58  global alarm_indicators weekly_average_temperature weekly_total_rainfall weekly_average_humidity
59
60  *Window size for measuring mean alarm indicator inluding current week (1=only current week, 2=current week +
61  scalar alarm_window=3
62
63  *Alarm threshold: is the value from which the "alarm signal" can be declared.  Alarm signal can occur when th
64  *For example, for a given value if prob_alarm=0.2, if the outbreak probability was > 0.2 thhen we say we have
65  scalar alarm_threshold=0.2
```

Line number: 58

Choice of alarm indicator(s)

global alarm_indicators weekly_average_temperature weekly_total_rainfall weekly_average_humidity

- Alarm indicators are defined as an alarm that can predict a forthcoming outbreak.
- Here you can choose which alarm indicators you want to test for predictive capacity.
- Enter the desired alarm indicator(s) according to the name of the column in your surveillance dataset: you may include an unlimited number of alarm indicators in this command.
- Missing data will negatively affect the results so be sure that you have a complete dataset before running any or multiple alarm indicators.
- Do not alter the "alarm_indicators" text.
- Remember to keep a space between each alarm indicator you enter, e.g. weekly_average_temperature weekly_total_rainfall.

- It is important that the name of the alarm indicator you enter in this option is EXACTLY the same as the variable name in the surveillance data (i.e. the same text you find in the column title in your surveillance data).
- *This corresponds to* **step 4** *in the annex.*

Line number: 61

Choice of window size for the alarm indicators

```
scalar alarm_window= 3
```

- The values of each alarm indicator will be recalculated by Stata to produce an average over a given time period. Here you can alter this time period.
- Enter the desired choice of window size (number of weeks) from which the mean value of the alarm indicator can be calculated.
- For example, entering a value of 3 for this option suggests that you are calculating the mean value (average) of the alarm indicator (e.g. temperature) of the current week and two previous weeks, consecutively. A minimum value of "1" means you are measuring the alarm indicator unit of the current week!
- *This corresponds to* **step 4** *in the annex.*

Line number: 65

Define the alarm threshold

```
scalar alarm_threshold= 0.15
```

- Here you decide what threshold is used to signal an alarm. If the threshold is too high, you will not have any alarms and, therefore, will not predict any outbreaks. If the threshold is too low, you will have many alarms, detect all outbreaks but will also have many false alarms. You need to experiment by changing the threshold to get a good balance between a low number of false alarms (positive predictive value or PPV) and high outbreak prediction rate (sensitivity). A higher PPV indicates lower false positive alarms!
- We suggest you use a value between 0.05–0.2.
- *This corresponds to* **step 13** *in the annex.*

2.1.5 Settings for running graphics

Figure 8. Graphical Settings

```
68   ********************************* GRAPH SETTINGS *********************
69
70   *Specifies the district for which the graphs (three graphs; run-in, evaluation
71   *If graphs for "All" districts is desired, then no entery is required i.e. keep
72   global graph_per_district 5
73
```

Line number: 72

Selecting your graphical outputs

global graph_per_district= *15*

- Here you can generate a graph for a specific district, many districts or all districts.
- Enter the desired district code(s) for the graph presentation of the outcome. For example, graph_per_district =15 will generate the analysis graphs (run-in, evaluation and both together) for district 15 only.
- You can enter more than one code in this line by keeping a space between each code. For example, graph_per_district= 15 16 19 will generate graphs for districts 15, 16 and 19.
- Leaving this option "blank" will generate a graph for the entire data.
- *This corresponds to* **step 14** *in the annex.*

2.1.6 Advanced settings

Figure 9. Advanced Settings

```
75  ********************************ADVANCE SETTINGS!!*******************
76
77  * Estimate the regression based on seasonal relations. You can choose the length
78  * Season_length=4 means that the first regression is based on the first 4 weeks
79  scalar season_length=52
80
81  *The multiplier of the standard deviation to vary the endemic channel within the
82  scalar z_outbreak=1.25
83
84  *Outbreak window - mean of outbreak weeks are summarized within the outbreak win
85  *Window size (1 is one week, 2 is two weeks etc.):
86  scalar outbreak_window=4
87
88  *Choice of distance and window size: specifying distance between current week an
89  *E.g. for week 10 and distance=2&outbreak_window=4 it would start measuring from
90  scalar prediction_distance=2
91
92  *This refers to the cut-off value to define the outbreak signal.
93  *The given cut-off value will convert the proportions (i.e. number of outbreaks
94  *For example, if the derived proportion=0.75 (i.e. 3 outbreak weeks during outbr
95  scalar outbreak_threshold=0.75
96
97  *Spline option yes=1 or no=0. Spline is used to capture non-monotonic relations
98  scalar spline=0
99
```

In this section you can further modify the do-file using additional settings. However, it is not necessary to run the do-file. Only attempt this section if you are comfortable with the changes you made above.

Line number: 79

Choice of the seasonal variation

scalar season_length= *52*

- If you believe that some alarm indicators are better predictors at different times of the year, you can allow for this by dividing the year into different periods.
- For example, season_length=4 means that the first alarm indicator analysis is based on the first 4 weeks of the year, the second is based on week 5–8, the third is week 9–12, etc.
- *This corresponds to* **step 8** *in the annex.*

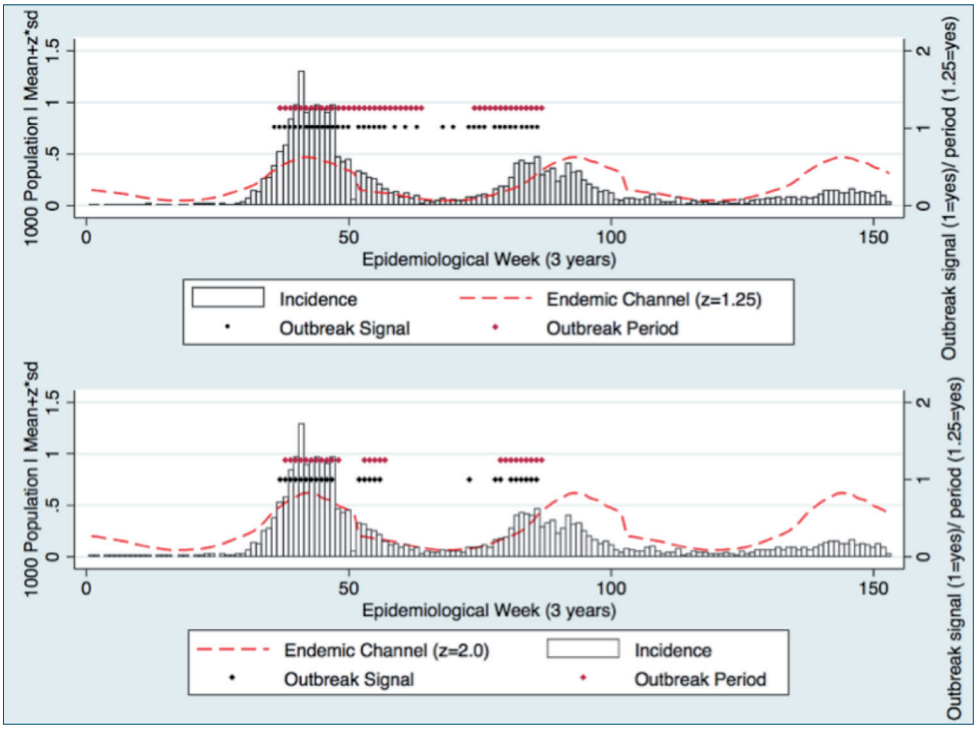
Line number: 82

Changing the endemic channel (the outbreak threshold)

```
scalar z_outbreak= 1.25
```

- Here you can change the standard deviation which increases or decreases the outbreak threshold. By increasing this threshold you will define fewer outbreaks, and by decreasing this you will increase the number of recorded outbreaks (Figure 10).
- Enter the desired value to define the multiplying value by the standard deviation (SD), for example, z=1 is the same SD, z=1.5 is one and half times the SD, z=2 is two times the SD etc.
- *This corresponds to* **step 2** *in the annex.*

Figure 10. Modelling the Endemic Channel



Line number: 86

Choose the denominator required to define the number of outbreak weeks

```
scalar outbreak_window= 4
```

- In a previous step (line 52 in your do-file), you defined the outbreak period. Here, in this step, you can choose a suitable window size (or denominator) for calculating the proportion from these outbreak periods. For example, if you choose a window size of 4, then Stata will take the "sum of values" of each consecutive 4 outbreak periods and divide it by 4 to generate a proportion – this proportion is then needed in the next steps to formulate the outbreak signal.
- Decreasing this window size can increase the sensitivity of predicting an outbreak. At the same time, you may have to choose to increase this window size when you have some missing records in your data.
- The choice of your window size depends on the evaluation criteria that give the best outbreak predictions, see step 14 in the annex.
- *This corresponds to* **step 5** *in the annex.*

Line number: 90

Choice of distance (in weeks) from where the outbreak window is calculated

```
scalar outbreak_window= 4
```

- Enter the desired choice of distance between current week and target week to predict an outbreak signal.
- For example, at week 10, if prediction_distance=2 and outbreak_window=4 (previous option), this will mean that you start counting from week 12 up to week 15 (4 weeks) to predict an outbreak signal.
- *This corresponds to* **step 5** *in the annex.*

Line number: 95

Defining the cut-off point of the outbreak signal

```
scalar outbreak_threshold= 0.5
```

- In a previous step (line 86 in your do-file), you computed the "proportions" from the outbreak periods. The proportion is a value between "zero" and "one" and in order to process this value to generate an outbreak signal, it needs to be converted into either "zero" (which means no outbreak) or "one" (which means an outbreak).

- Here, in this step, you can choose a cut-off value to define this outbreak signal. For example, if you choose a cut-off value of 0.5, then every proportion value above this cut-off value (e.g. 0.7> 0.5) will be given the code "one" (indicating that there is an outbreak), and any proportion value less than this cut-off (e.g. 0.3< 0.5) will be given the code "zero" (indicating no outbreak). This way you have defined your outbreak outcome (a binary variable (0, 1), which is the dependent variable needed for the logistic regression.
- The choice of a relevant value depends on the evaluation criteria that give the best prediction of outcome, see step 14 in the annex.
- *This corresponds to* **step 7** *in the annex.*

Line number: 98

Choice of the spline application

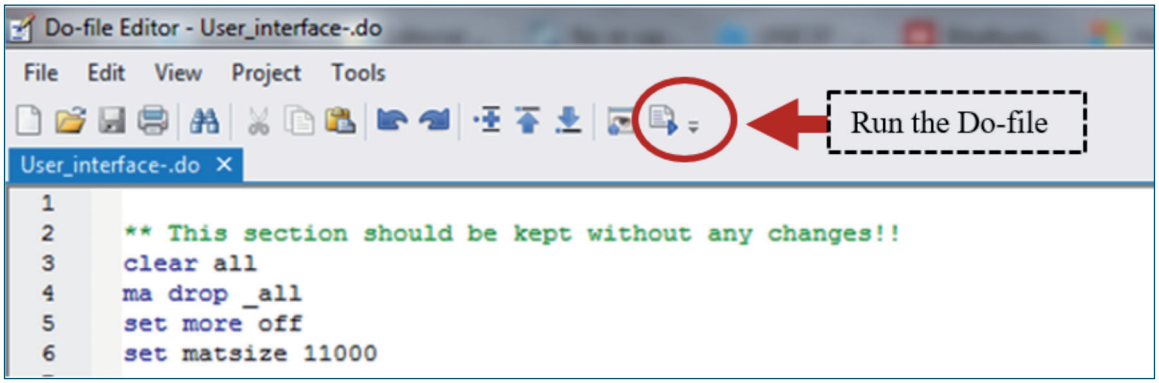
```
scalar spline= 1
```

- Spline is a function that allows Stata to treat a certain type of relationship between alarm indicators and dengue cases.
- Without spline, Stata will assume linear (non-monotonic) relationships, i.e. as temperature increases, dengue cases increase.
- With spline, Stata will assume monotonic relationships, i.e. as temperature increases, dengue cases go down. Here you can choose between two options.
- Enter the desired spline option, i.e. code 0=No (non-monotonic relation) or, 1=YES (monotonic relation).
- You can see how spline affects your results by first running without spline and then with spline.
- *This corresponds to* **step 6** *in the annex.*

2.1.7 Run the do-file

Once you have completed the changes, please run the do-file!

Figure 11. Run the Do-File



2.1.8 Interpreting the results of the prospective data

When the do-file has been run there will be two tables of results displayed in Stata. The first table is shown in Figure 12.

Figure 12. Absolute Values in the Output Table

District	Weeks	Outbreak Weeks	Outbreak Periods	Alarms	Correct Alarms	False Alarms	Missed Outbreaks	No Alarms, No Outbreaks
12	51	32	34	12	10	2	25	14
15	51	34	35	17	16	1	18	16
Total	**102**	**66**	**69**	**29**	**26**	**3**	**43**	**30**

The definitions of each of the column headings are :

1. **Outbreak weeks:** a week where the number of cases is above the endemic channel (z value).
2. **Outbreak periods:** consecutive outbreak weeks (for our definition, this is an outbreak).
3. **Alarms:** total number of alarms.
4. **Correct alarms:** alarm correctly predicts the outbreak.
5. **False alarms:** alarm falsely predicts the outbreak (false positive).
6. **Missed outbreaks:** alarm that could predict the outbreak if the alarm threshold was lower.
7. **No alarm, no outbreak (negative correct alarms):** no alarms when there is no outbreak.
8. **Defined_outbreaks:** this is the total number of outbreaks as defined by the given "outbreak threshold" (cut-off) value.

Together, these absolute numbers show the results of the retrospective phase and the success/ failure of the parameters (alarm threshold, z value, etc.) that were used to analyse the dataset.

The second table describes the proportion of successfully detected outbreaks (sensitivity) and the proportion of false alarms (PPV) (Figure 13).

Definitions (tests of validity):

- Sensitivity: the percentage of outbreaks correctly predicted by alarms. This should be as close to 100% as possible. Minimum value of 50%.
- PPV: the percentage of alarms that incorrectly predicted outbreaks (higher PPV indicates lower false positive alarms). This should be as close to 100% as possible. Minimum of 50%.
- To increase the number of correct alarms and decrease the number of false alarms, we suggest that you alter the z value and alarm threshold, then, if necessary, alter the remaining parameters as desired. Then rerun the do-file.

Figure 13. Sensitivity and PPV

Summary statistics: sum
by categories of: district

district	sensit~y	PPV
15	.52	.9
Total	.52	.9

2.1.9 Interpreting the results from an operational perspective

If sensitivity is 60%, then for every 10 outbreaks, the early warning system will detect 6 outbreaks correctly. It will miss 4 outbreaks.

If PPV is 70%, 7 out of 10 alarms will be correct, 3 out of 10 alarms will be false. That means that 3 times out of 10, resources will be incorrectly mobilized, i.e. wasted.

Consider this: below, you will build the early warning system using your retrospective data. The results above are based on the presence of 1 alarm signal before an outbreak. The early warning system below uses a total of 3 alarm signals to better warn of outbreaks. Therefore, in the presence of 2 or 3 (or more) alarms, the chance of outbreak detection increases and the wasting of resources decreases with each alarm.

Chapter 3
Using the early warning and response system

3.1 Prospective early warning system

In this chapter we discuss how to build your prospective early warning system using results from your retrospective analyses in the previous chapters.

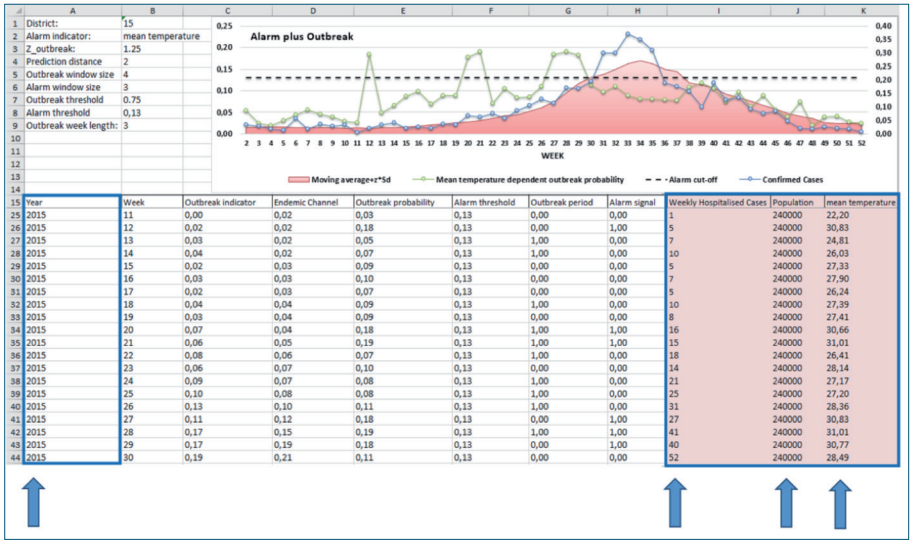
You have chosen which alarm indicators are best predictors of outbreaks in your district/ country. Now we need to apply these within an early warning system that will allow you to detect outbreaks in real-time.

By the end of this chapter, you will be familiar with the Excel outputs, data entry and interpretation of alarms.

To generate the early warning system in Excel, you need to enter a '1' in Line 34 of the do-file. As described above in section 2.1.2 "Retrospective analysis or prospective early warning", selecting this option will build the Excel-based early warning system (Figure 14) where you can enter data in real-time to predict outbreaks in the future.

NB: Before running the analysis, you will need to change your surveillance workbook file name. Do this by adding the District Number to the surveillance workbook file name. For example, if you want to run the analysis for District "15" then you will need to save the surveillance workbook Excel file as: "surveillance workbook_15.xlsx". If you want to analyse multiple districts at the same time, copy/paste the workbook and add the additional district numbers. For example, to analyse district 15 and 30, you will need two workbooks, one called "surveillance workbook_15.xlsx" and the other called "surveillance workbook_30.xlsx". This step is only necessary when generating the workbook excel file.

Figure 14. Prospective Data Entry and Interpretation*



1	District:	15
2	Alarm indicator:	mean temperature
3	Z_outbreak:	1.25
4	Prediction distance	2
5	Outbreak window size	4
6	Alarm window size	3
7	Outbreak threshold	0.75
8	Alarm threshold	0,13
9	Outbreak week length:	3

	Year	Week	Outbreak indicator	Endemic Channel	Outbreak probability	Alarm threshold	Outbreak period	Alarm signal	Weekly Hospitalised Cases	Population	mean temperature
25	2015	11	0,00	0,02	0,03	0,13	0,00	0,00	1	240000	22,20
26	2015	12	0,02	0,02	0,18	0,13	0,00	1,00	5	240000	30,83
27	2015	13	0,03	0,02	0,05	0,13	1,00	0,00	7	240000	24,81
28	2015	14	0,04	0,02	0,07	0,13	1,00	0,00	10	240000	26,03
29	2015	15	0,02	0,03	0,09	0,13	0,00	0,00	5	240000	27,33
30	2015	16	0,03	0,03	0,10	0,13	0,00	0,00	7	240000	27,90
31	2015	17	0,02	0,03	0,07	0,13	0,00	0,00	5	240000	26,24
32	2015	18	0,04	0,04	0,09	0,13	1,00	0,00	10	240000	27,39
33	2015	19	0,03	0,04	0,09	0,13	0,00	0,00	8	240000	27,41
34	2015	20	0,07	0,04	0,18	0,13	1,00	1,00	16	240000	30,66
35	2015	21	0,06	0,05	0,19	0,13	1,00	1,00	15	240000	31,01
36	2015	22	0,08	0,06	0,07	0,13	1,00	0,00	18	240000	26,41
37	2015	23	0,06	0,07	0,10	0,13	0,00	0,00	14	240000	28,14
38	2015	24	0,09	0,07	0,08	0,13	1,00	0,00	21	240000	27,17
39	2015	25	0,10	0,08	0,08	0,13	1,00	0,00	25	240000	27,20
40	2015	26	0,13	0,10	0,11	0,13	1,00	0,00	31	240000	28,36
41	2015	27	0,11	0,12	0,18	0,13	0,00	1,00	27	240000	30,83
42	2015	28	0,17	0,15	0,19	0,13	1,00	1,00	41	240000	31,01
43	2015	29	0,17	0,19	0,18	0,13	0,00	1,00	40	240000	30,77
44	2015	30	0,19	0,21	0,11	0,13	0,00	0,00	52	240000	28,49

3.1.1 Data entry

Once the file is built, you can enter data on a weekly basis. You will need to enter data into the empty columns (Figure 14, selected in blue) under the following headings: year, number of cases, population, and alarm variables such as mean temperature. In Cell B2 at the top left-hand corner of the spreadsheet, you will see the list of alarm indicators that have been built into the spreadsheet. You will need to enter data for all the alarm indicators listed here.

Automated graphs:

- Once the user enters the required information, the graphs will automatically be updated in the same sheet.

Results interpretation:

- Reflected in the top left-hand corner, you can see all the settings that were chosen in Stata. If you want to change these settings, you must do it in the Stata do-file (user_interface.do). Then rerun the do-file to build a new Excel workbook. Remember to define generating_surveillance_workbook=1 to repopulate your Excel file.

Let us use Figure 14 as a worked example to explain each of the settings and where they can be found in Stata.

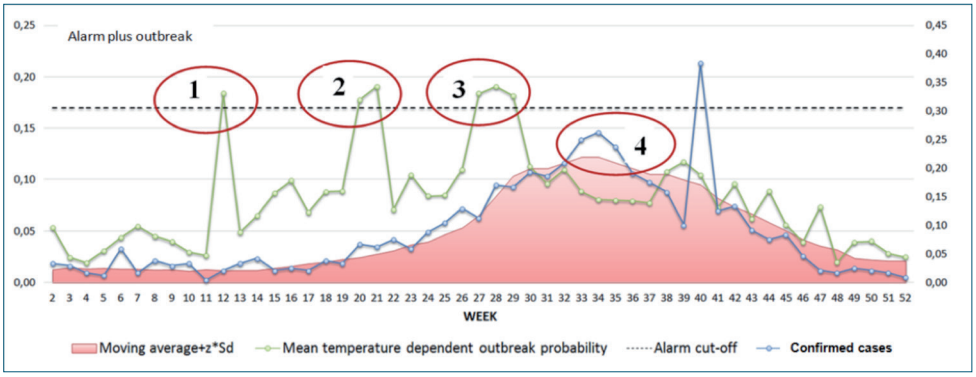
- This analysis is related to district 15.
- The alarm indicator is "mean temperature". Therefore, you must enter weekly mean temperature data into the column.
- The endemic channel (outbreak threshold) was chosen as 1.25 (this is the multiplier of the standard deviation of the moving average).
- The prediction distance=2: this is the distance between the current week and the predicted week.
- The outbreak window size=4: this represents the denominator used for computing the proportion from the outbreak periods that are later processed to formulate the outbreak signal.
- The alarm threshold=0.13: this is the threshold that is used to define the alarm signal (if the outbreak probability were ≥ 0.13 we would have alarm signal=1).
- In the "Alarm signal" column, where the entry is '1', this indicates that the alarm signal has correctly predicted a future outbreak.

3.1.2 Interpretation of staged response

Example of how to interpret the results (see Figure 15):

- No response: 1 positive alarm week;
- Initial response: 2 positive alarm weeks;
- Early response: 3 positive alarm weeks;
- Late/emergency response: 3 positive outbreak weeks.

Figure 15. Results Interpretation: An Illustrated Example



1 = **No response**
2 = **Initial response**
3 = **Early response**
4 = **Late/emergency response**

Annex 1.
Technical guide

1. Introduction

This technical guide explains the scientific rationale behind the Early Warning and Response System (EWARS) for Dengue Outbreaks: Operational guide.

This project has focused on developing a validated Stata-based model that can enable the prediction of 'out of control' dengue cases (outbreaks) as defined by dengue incidence (probable/hospitalized cases). The early warning system detects changes in the alarm indicators (entomological, meteorological and epidemiological) to predict dengue outbreaks.

The purpose of this Technical Guide is to provide an overview of the applied method and its rationale. Additionally, it will present block diagrams to describe the applied method with further details of each step. This will assist the user to follow the Guide and link to the corresponding step number. It will also discuss general requirements for the Stata program and potential errors while using the available do-files.

2. Methodology

The general methodological concept of this analysis follows two major phases that are summarized below.

2.1 Retrospective phase

This phase uses retrospective surveillance data to create two datasets: (i) run-in data, used to develop the prediction model; and (ii) evaluation data, used to evaluate the derived parameters from the prediction model.

2.1.1 Run-in data

This dataset uses past records to estimate/ calibrate the model parameters of the relationship between the outbreak indicator and alarm indicator(s). This parameterization is then tested during the evaluation process (step 2, below) and applied by the user to predict an outbreak. These data include information on the year, week and district, outbreak indicator (probable, confirmed cases, or other forms of outbreak indicators), and alarm indicator(s), such as the weekly mean temperature, sum of rainfall, mean humidity and probable cases.

2.1.2 Evaluation data

This dataset is used to: (a) evaluate the prediction model; and (b) provide summary statistics that are used to build the prospective early warning system.

2.2 Prospective early warning phase (i.e. surveillance workbook Excel file)

The populated Excel file of results (final parameters) allows the user to enter prospective information to estimate the probability of an outbreak for a forthcoming period. This is simply derived by inserting weekly data of outbreak and alarm indicator(s) for the district(s) of interest.

Rationale for the approach

The final model derived for estimating the probability of an outbreak is generated via systematic steps (creating/evaluating parameters) by assessing the association between the level in the alarm indicators (i.e. mean rainfall, mean temperature, mean of humidity etc.) and the dengue outbreaks. Logistic regression is used to assess this association. This regression model processes the computed proportions of the outcome (computed via a cut-off value, given by the user), to predict the probability of dengue outbreaks for a forthcoming period. Throughout these steps, the user is able to define relevant measures such as the size of the endemic channel, the outbreak duration, the alarm threshold and the prediction distance. Descriptions of these measures can be found in the following chapter.

3. Structural design (Stata and Excel)

The block diagram in Figure 16 outlines 16 steps in the overall process distributed across two different phases. Phase I (retrospective) is divided into: (i) the run-in data; and (ii) the evaluation data. Phase II (prospective) is the final analysis using summary statistics to populate an Excel-based early warning system that can be used in real-time to detect future dengue outbreaks. A summary of each step is presented below.

3.1 Phase I (retrospective)

3.1.1 Using the run-in data to create the prediction model (part one)

Step 1

In this step, the original data are divided into run-in and evaluation data. The user determines this by entering the period (cut-off), in year-week, when the run-in data end and the evaluation data begin. A minimum of two years' data are required for the "run-in data" though more than two years' data are recommended.

Step 2

This step refers to the endemic channel and is represented by the following equation:

$$\text{Endemic channel} = (Z*SD) + \text{moving average}$$

"Z" is a multiplier value of the SD provided by the user to vary the breadth of the endemic channel. This part will assist in declaring the "out of control" status. For instance, a value of Z=1.5 would increase the breadth one and a half times the expected "normal" range of the number of dengue cases. The moving average is the mean number of dengue cases within the expected "normal" or seasonal range, calculated for a window size of three preceding and three succeeding weeks from the point (week) of measure.

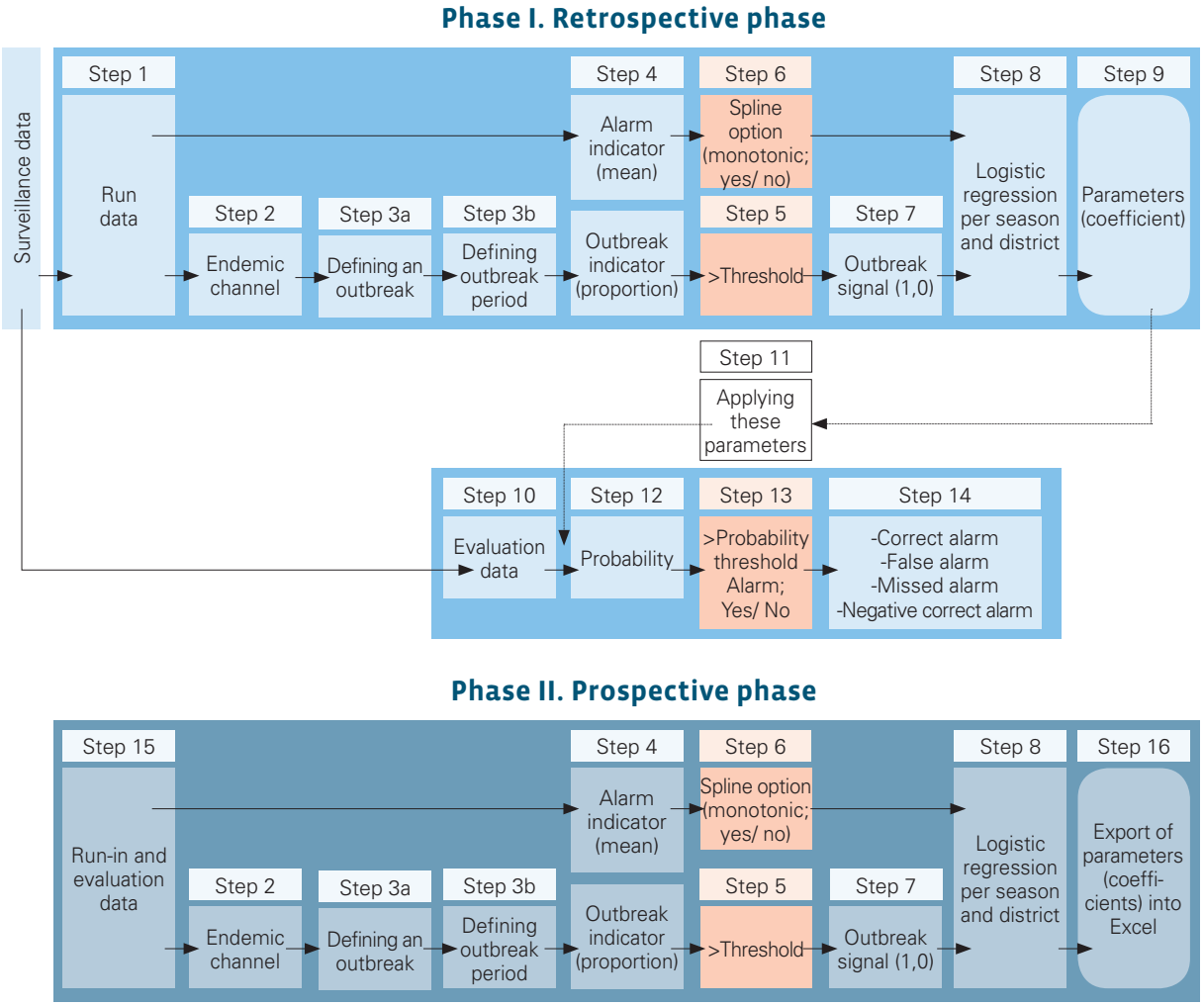
Step 3a

In this step, an "outbreak" is defined by the proportion of number of incident cases in relation to the annual population for a corresponding district. The user is given the option to choose which outbreak indicator to use, for example, probable or hospitalized cases.

Step 3b

In this step, the user may choose the desired length of an outbreak period. The user can define the length of the outbreak period by determining the number of consecutive weeks (1, 2, 3, etc.) that dengue cases must exceed the endemic channel. For example, for outbreak_week_length=3, we can say that our outbreak period is defined when the dengue cases continue to exceed the endemic channel value for "3" consecutive weeks.

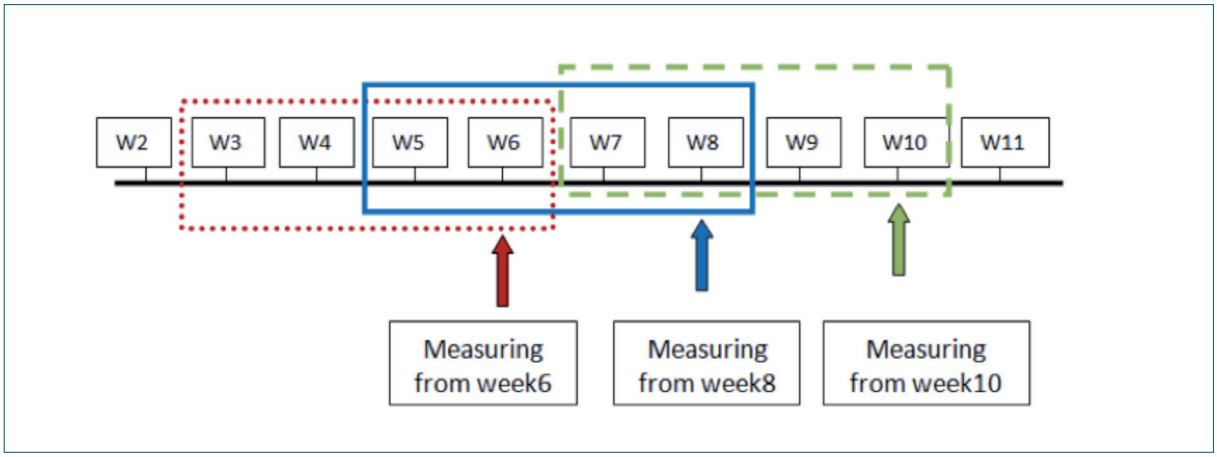
Figure 16. Block Diagram of the Retrospective/ Prospective Phases



Step 4

This step illustrates how data from the alarm indicator is used in relation to the alarm window (Figure 17). We measure the mean of each alarm indicator (mean temperature, weekly rainfall, etc.) during the "preceding" pre-defined alarm window size (e.g. choosing alarm window size of 4, this step will measure the mean of each of these alarm indicators during the last four consecutive weeks including the week we are measuring from).

Figure 17. Illustrated Process of Alarm Window Size



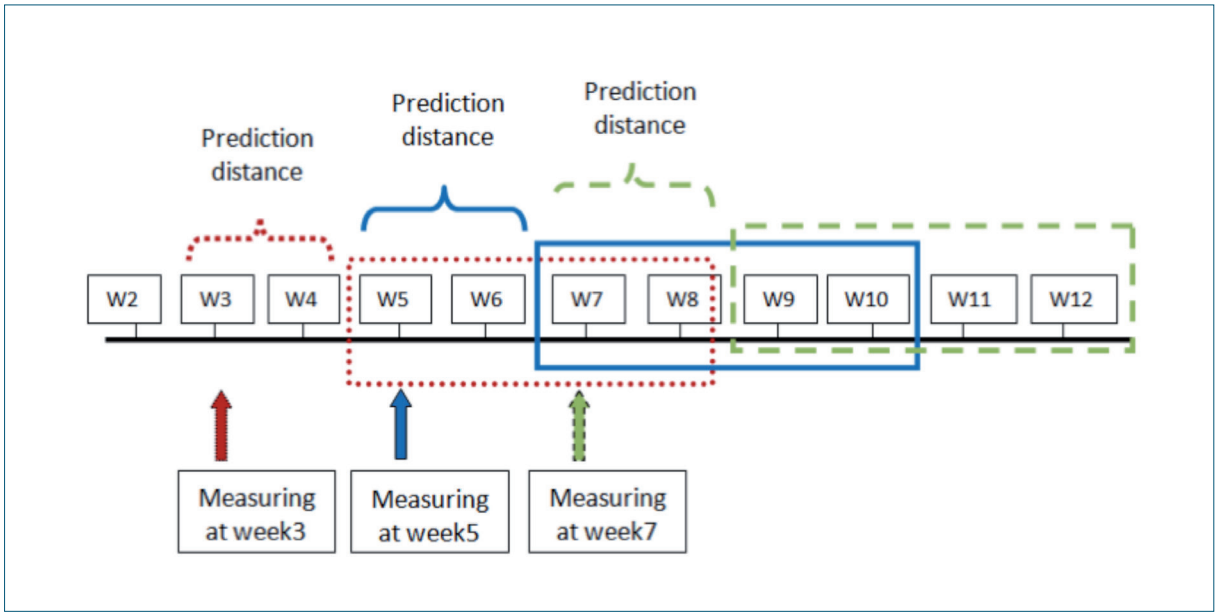
Step 5

This step illustrates how data from the outbreak period are used with the window size to define an outbreak (Figure 18).

Here, the formulated outbreak period (from step 3b) is converted into proportions by dividing the "sum of values" from the outbreak periods by a given window size (or denominator) which generates values that lie between "zero" and "one" (proportion). The user usually defines the window size during the calibration stage which defines these "proportions", which will in later steps be used to form the outbreak outcome (using the outbreak threshold).

For example, if the user chooses a window size of 4, then Stata will take the "sum of values" of each consecutive 4 outbreak periods and divide it by 4 to generate a proportion. Figure 18 below illustrates this process by also describing how other information (also defined by the user) can be used in this step. The prediction distance is another factor used in this process that defines the point where computing the proportion should begin in the data.

Figure 18. Illustration of Prediction Distances



Step 6

The spline option handles a monotonic relationship between the alarm indicator(s) and outbreak indicator. A monotonic relationship occurs when there is an increase in the mean alarm indicator during a particular period with a decreased number of dengue cases (outbreak) for the same period. Applying this option (yes=1 (monotonic)/no=0 (non-monotonic)) is decided during the evaluation stage (section 1.2.3 Choice of the spline application; line 35 in the do-file).

The user can observe the performance (evaluation table) of the prediction model when the spline is selected or removed. The method of checking the performance of the prediction model is illustrated in step 14 below.

Step 7

In a previous step (step 5), proportions from the outbreak periods were computed using a pre-defined window size and prediction distance. In this step, the outbreak signal is formulated by converting the derived proportions into a binary variable (0=no outbreak or, 1=outbreak), which is necessary for running the logistic regression in this next step.

In this step (7), the user chooses a suitable cut-off point (threshold) which defines an outbreak signal (i.e. code “1”) or not (i.e. code “0”). For example, if you choose a threshold value of 0.5, then any proportion above this threshold (e.g. 0.7> 0.5) will be given the code “one” (an outbreak signal) and any proportion value less than this threshold (e.g. 0.3< 0.5) will be given the code “zero” (not an outbreak signal). This will result in a binary “outbreak” outcome variable that is ready to enter the logistic regression process.

Step 8

In this step, we assess the association between the binary outcome of the outbreak signal (0, 1) and the alarm indicators using logistic regression. This relationship will be influenced by the "season_length" input. For example, if season_length=4 weeks, the logistic regression model will generate a coefficient for this period length alone. Then the next coefficient will be specific to the following consecutive 4 weeks and so forth.

Step 9

The coefficients generated from this regression model will be stored/ applied during the evaluation stage.

3.1.2 Using the evaluation data to assess how the derived parameters from the run-in data would predict an outbreak (part two)

Step 10

Refers to the second part after splitting the surveillance data, which we call the "evaluation data" (the first part of the original data was used in "Step one", i.e. the run-in data). The user determines this by entering the period, in 'year-week' for when the run-in data ends and the evaluation data begins. A minimum of two years' data records is required for each dataset.

Step 11

In this step, we are evaluating the coefficients that were initially generated by the run-in data. This evaluation is performed by applying these coefficients to the evaluation data to observe the performance of the prediction model in detecting outbreaks.

Step 12

The derived function from the logistic regression step 8 and its parameters is used to estimate the probability of an outbreak occurrence.

Step 13

The generated probability from step 12 will require a threshold to determine an alarm signal. For instance, for a probability of 0.4, with a threshold of 0.3 (i.e. the probability ≥ threshold), this record is said to be an alarm signal. If, however, the probability were less than the threshold, then it is not an alarm signal.

Step 14

In order to ensure that the given threshold (in step 13) is reliable to predict an outbreak, the Stata program will further present four evaluation criteria to assess the choice given in step 13. These criteria are:

1. correct alarm: probability ≥ threshold with a true outbreak (for a target period);
2. false alarm: probability ≥ threshold with not true outbreak (for a target period);
3. missed outbreak: probability ≤ threshold with a true outbreak (for a target period);
4. no alarm, no outbreak: probability ≤ threshold with no true outbreak (for a target period).

An optimal threshold (option in step 13) would lead to an increased number of correct alarms and negative correct alarms but a decreased number of false alarms and missed alarms. By observing these results (which will be displayed on the screen of the Stata program – output), the user can alter the threshold in step 13, accordingly.

3.2 Phase II (prospective surveillance workbook)

Step 15

In this step, the program will run through the entire dataset (i.e. both run-in + evaluation) to populate the output Excel file (surveillance workbook) that includes information on the probability formula, model parameters and details on the applied settings from the evaluation period. Once the user has entered real-time prospective data into the empty columns, the file will automatically estimate and graph the probability of an outbreak.

Steps 2–8 (are fixed in this phase)

Settings from these steps will be fixed according to the settings derived during the evaluation stage, i.e. we follow the same settings that we found (via the evaluation stage) to provide best prediction.

Step 16

In this final step, the Stata program will populate the surveillance workbook. The user will input data on year, week, outbreak indicator and alarm indicator(s) for a corresponding district into this Excel sheet, which can automatically calculate the probability of an outbreak and produce an instant graphical presentation.

4. Potential errors

Potential errors and the corresponding sources/solutions.

4.1 Problem related to Step 2: Few Data (error code r 2000)

An error could occur if the "run-in data" consist of total records of less than a two-year period (i.e. the SD will have no values – missing)!

Solution: ensure a minimum number of records covering a two-year period for the run-in data (i.e. defining the year_week period, accordingly).

Figure 19. Example Error Message: "No Observations"

```
. ******************************** EXPORTING RESULT SETTINGS**************
.
.
. run "Dengue_Analysis_phase IIv5.do"
no observations
r(2000);

end of do-file
```

4.2 Problem related to Step 7: Convergence not achieved (error code r 430)

Outbreak "threshold" value generates the binary variable, i.e. "0" and "1". Choosing too high/ too low outbreak threshold value could result in having the binary variable reading only "0" or only "1" which can restrict the regression process.

Solution: if such error occurs, adjust the outbreak threshold value.

Figure 20. Example Error Message: "Convergence not Achieved"

```
. ******************************** EXPORTING RESULT SETTINGS**************
.
.
. run "Dengue_Analysis_phase IIv5.do"
convergence not achieved
convergence not achieved
r(430);

end of do-file
```

4.3 Problem related to Step 8: too low/ high values of "season length" or "outbreak period" (error code r 2000)

Another potential error occurs in this step when the value for the season_length is too small or the outbreak_period_length is too large (e.g. =15 weeks).

Solution: increase the season_length (value) or reduce the given value for the outbreak_week_length.

Figure 21. Example Error Message: "r2000"

```
. ********************************* EXPORTING RESULT SETTINGS***************
.
.
. run "Dengue_Analysis_phase IIv5.do"
Regression for district 15 has failed
Regression for district 15 has failed
Regression for district 15 has failed
r(2000);

end of do-file
```

4.4 Problem related to Step 16: open/ active excel file before running the analysis (error code r 603)

An error could arise if the output Excel file is open (or active).

Solution: the user needs to make sure that this file is closed before running the analysis (the do-file).

Figure 22. Example Error Message: "File [...] Not Saved"

```
. ********************************* EXPORTING RESULT SETTINGS***************
.
.
. run "Dengue_Analysis_phase IIv5.do"
Regression for district 15
file Surveillance workbook_15.xlsx could not be saved
r(603);

end of do-file
```

4.5 Problem related to Step 8: too small sample size (error code r 198)

An error could arise if there are missing records in the applied variable.

Solution: the user needs to complete the data records or, remove the variable from the analysis (e.g. alarm indicator) that has high missing records.

Figure 23. Example Error Message: "Sample Size Too Small"

```
. ******************************** EXPORTING RESULT SETTINGS**************
.
.
. run "Dengue_Analysis_phase IIv5.do"
Sample size too small for this many knots.
invalid syntax
r(198);

end of do-file
```

4.6 General error: wrong district code (error code r 111)

An error could arise if the user selects to run the analysis for a district that does not exist in the data.

Solution: ensure that the district code exists and has records in the file.

Figure 24. Example Error Message: "Variable_000001 Not Found"

```
. ******************************** EXPORTING RESULT SETTINGS**************
.
.
. run "Dengue_Analysis_phase IIv5.do"
variable __000001 not found
r(111);

end of do-file
```

NOTES

The Special Programme for Research and Training in Tropical Diseases (TDR) is a global programme of scientific collaboration established in 1974. Its focus is research into neglected diseases of the poor, with the goal of improving existing approaches and developing new ways to prevent, diagnose, treat and control these diseases. TDR is sponsored by the following organizations:

unicef · UNDP · THE WORLD BANK IBRD • IDA | WORLD BANK GROUP · World Health Organization

TDR For research on diseases of poverty
UNICEF • UNDP • World Bank • WHO

TDR/World Health Organization
20, Avenue Appia
1211 Geneva 27
Switzerland

Fax: (+41) 22 791-4854
tdr@who.int
www.who.int/tdr

ISBN 978 92 4 151205 3

9 789241 512053